Supplementary Materials: “Confidence Intervals for Sparse Penalized Regression with Random Designs”

A Basics in variational inequalities and the normal manifold

The tangent cone to S at x is defined as

$$T_S(x) = \{w \in \mathbb{R}^n | \ \exists \{x_k\} \subset S \text{ and } \{\tau_k\} \subset \mathbb{R} \text{ such that } x_k \to x, \tau_k \to 0, \text{ and } (x_k - x)/\tau_k \to w\}.$$

The inner product of any element in $T_S(x)$ and any element in the normal cone $N_S(x)$ is nonpositive.

Consider a problem of minimizing a objective function $F : \mathbb{R}^n \to \mathbb{R}$ over a closed and convex feasible set S. The well-known first-order necessary condition is that, if $x^* \in S$ is a local solution to this minimization problem and F is differentiable at x^*, then the following variational inequality holds for x^*:

$$0 \in \nabla F(x^*) + N_S(x^*).$$

If the set S is a polyhedral convex set, then the Euclidean projector Π_S is a piecewise affine function on $\mathbb{R}^n$, that coincides with an affine function on each of finitely many n-dimensional polyhedral convex sets. This family of sets is called the normal manifold (Robinson, 1995) of S, and each set in this family is called an n-cell. The union of all n-cells in the normal manifold is $\mathbb{R}^n$. Faces of the n-cells are called cells, and the relative interiors of all cells form a partition of $\mathbb{R}^n$. More details can be found in Facchinei and Pang (2003) and Robinson (1992, 1995).

B-differentiability is related to directional differentiability, and it is stronger than directional differentiability. If $df(x_0)$ is the B-derivative of a function $f : \mathbb{R}^n \to \mathbb{R}^m$ at x_0, then for each direction $h \in \mathbb{R}^n$, $df(x_0)(h)$ is exactly the directional derivative of f at x_0. In addition, B-differentiability requires $df(x_0)(\cdot)$ to be a first order approximation of $f(x_0 + \cdot)$ uniformly in all directions.

Cell	Defining constraints	Critical cone	Defining constraints
C_i^0	$t_i = 0, \quad \beta_i = 0$	K_i^0	$t_i - \beta_i \geqslant 0, \quad t_i + \beta_i \geqslant 0$
C_i^1	$t_i = \beta_i, \quad t_i \geqslant 0$	K_i^1	$t_i - \beta_i \geqslant 0$
C_i^2	$t_i = -\beta_i, \quad t_i \geqslant 0$	K_i^2	$t_i + \beta_i \geqslant 0$
C_i^3	$t_i = \beta_i, \quad t_i \leqslant 0$	K_i^3	$t_i = -\beta_i, \quad t_i \geqslant 0$
C_i^4	$t_i = -\beta_i, \quad t_i \leqslant 0$	K_i^4	$t_i = \beta_i, \quad t_i \geqslant 0$
C_i^5	$t_i - \beta_i \geqslant 0, \quad t_i + \beta_i \geqslant 0$	K_i^5	None
C_i^6	$t_i - \beta_i \geqslant 0, \quad t_i + \beta_i \leqslant 0$	K_i^6	$t_i = -\beta_i$
C_i^7	$t_i - \beta_i \leqslant 0, \quad t_i + \beta_i \leqslant 0$	K_i^7	$t_i = 0, \quad \beta_i = 0$
C_i^8	$t_i - \beta_i \leqslant 0, \quad t_i + \beta_i \geqslant 0$	K_i^8	$t_i = \beta_i$

Table 1: Cells in the normal manifold of S_i and the associated critical cones

	C_i^5	C_i^6	C_i^7	C_i^8
ψ_0	A_1	A_2	A_4	A_3
ψ_1	A_1	A_1	A_3	A_3
ψ_2	A_1	A_2	A_2	A_1
ψ_3	A_2	A_2	A_4	A_4
ψ_4	A_3	A_4	A_4	A_3
ψ_5	A_1	A_1	A_1	A_1
ψ_6	A_2	A_2	A_2	A_2
ψ_7	A_4	A_4	A_4	A_4
ψ_8	A_3	A_3	A_3	A_3

Table 2: Matrix representations of ψ_j for $j = 0, \cdots, 8$

B Proofs

Proof of Lemma 1. Without loss of generality, suppose $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ is a local optimal solution to (9). Since $P_{\lambda_i}(\cdot)$ is nondecreasing and m is positive, it is obvious that $\tilde{t}_i = |\tilde{\beta}_i|$ for all $i = 1, \cdots, p$. Denote the objective function in (1) by $g_1(\beta_0, \beta)$ and the objective function in (9) by $g_2(\beta_0, \beta, t)$. Then there exists a neighborhood $\mathcal{B}_1$ at $(\tilde{\beta}_0, \tilde{\beta})$ in $\mathbb{R}^{p+1}$, such that

$$g_2(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) \leqslant g_2(\beta_0, \beta, t) \quad \text{for each } (\beta_0, \beta) \in \mathcal{B}_1 \text{ and } t_i = |\beta_i|, \ i = 1 \cdots, p.$$

That is,

$$g_1(\tilde{\beta}_0, \tilde{\beta}) \leqslant g_1(\beta_0, \beta) \quad \text{for each } (\beta_0, \beta) \in \mathcal{B}_1.$$

Therefore, $(\tilde{\beta}_0, \tilde{\beta})$ is a local optimal solution to (1).

Piece	Defining constraints
E_i^0	$\lvert t_i - \beta_i\rvert \leqslant 1/g(N), \quad \lvert t_i + \beta_i\rvert \leqslant 1/g(N)$
E_i^1	$\lvert t_i - \beta_i\rvert \leqslant 1/g(N), \quad t_i + \beta_i > 1/g(N)$
E_i^2	$t_i - \beta_i > 1/g(N), \quad \lvert t_i + \beta_i\rvert \leqslant 1/g(N)$
E_i^3	$\lvert t_i - \beta_i\rvert \leqslant 1/g(N), \quad t_i + \beta_i < -1/g(N)$
E_i^4	$t_i - \beta_i < -1/g(N), \quad \lvert t_i + \beta_i\rvert \leqslant 1/g(N)$
E_i^5	$t_i - \beta_i > 1/g(N), \quad t_i + \beta_i > 1/g(N)$
E_i^6	$t_i - \beta_i > 1/g(N), \quad t_i + \beta_i < -1/g(N)$
E_i^7	$t_i - \beta_i < -1/g(N), \quad t_i + \beta_i < -1/g(N)$
E_i^8	$t_i - \beta_i < -1/g(N), \quad t_i + \beta_i > 1/g(N)$

Table 3: $E_i^0, \cdots, E_i^8$ in the plane (β_i, t_i)

Conversely, suppose $(\tilde{\beta}_0, \tilde{\beta})$ is a local optimal solution to (1). Then there exists a neighborhood $\mathcal{B}_2$ at $(\tilde{\beta}_0, \tilde{\beta})$ in $\mathbb{R}^{p+1}$, such that

$$g_1(\tilde{\beta}_0, \tilde{\beta}) \leqslant g_1(\beta_0, \beta) \quad \text{for each } (\beta_0, \beta) \in \mathcal{B}_2.$$

Let $\tilde{t}_i = |\tilde{\beta}_i|$ for all $i = 1, \cdots, p$, then we have

$$g_2(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) \leqslant g_2(\beta_0, \beta, t) \quad \text{for each } (\beta_0, \beta) \in \mathcal{B}_2 \text{ and } t_i = |\beta_i|,\ i = 1 \cdots, p.$$

Consequently,

$$g_2(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) \leqslant g_2(\beta_0, \beta, t) \quad \text{for each } (\beta_0, \beta) \in \mathcal{B}_2 \text{ and } t_i \geqslant |\beta_i|,\ i = 1 \cdots, p.$$

Thus, $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ is a local optimal solution to (9).

The second part of Lemma 1 is straightforward and we omit its proof.

□

Proof of Lemma 2. According to Assumption 3 and Lemma 1 we know that $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ is a local optimal solution to (9). We will prove that it is also a locally unique optimal solution by showing that L_K is a global homeomorphism.

From (12), we can write the normal and tangent cones to S at $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ as

$$N_S(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) = \{0\} \times N_{S_1}(\tilde{\beta}_1, \tilde{t}_1) \times \cdots \times N_{S_p}(\tilde{\beta}_p, \tilde{t}_p),$$

and

$$T_S(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) = \mathbb{R} \times T_{S_1}(\tilde{\beta}_1, \tilde{t}_1) \times \cdots \times T_{S_p}(\tilde{\beta}_p, \tilde{t}_p).$$

Let $\tilde{q}$ be as defined in Assumption 3, and let $\tilde{q}_0 = E[-2(Y - \tilde{\beta}_0 - \sum_{i=1}^{p} \tilde{\beta}_i X_i)]$. Since $-f_0(\tilde{\beta}_0, \tilde{\beta}, \tilde{t}) \in N_S(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$, we have

$$\tilde{q}_0 = 0 \text{ and } -(\tilde{q}_i - 2m\tilde{\beta}_i, P'_{\lambda_i}(\tilde{t}_i) + 2m\tilde{t}_i) \in N_{S_i}(\tilde{\beta}_i, \tilde{t}_i) \text{ for each } i = 1, \cdots, p. \tag{B.1}$$

If $\tilde{\beta}_i > 0$ for some $i = 1, \cdots, p$, from the definition of S_i and (B.1) we have

$$\tilde{q}_i - 2m\tilde{\beta}_i = -P'_{\lambda_i}(\tilde{t}_i) - 2m\tilde{t}_i.$$

That is

$$\tilde{q}_i = -P'_{\lambda_i}(\tilde{t}_i),$$

because $\tilde{t}_i = |\tilde{\beta}_i| = \tilde{\beta}_i$. Similarly, if $\tilde{\beta}_i < 0$, then

$$\tilde{q}_i = P'_{\lambda_i}(\tilde{t}_i);$$

if $\tilde{\beta}_i = 0$, then

$$|\tilde{q}_i| \leqslant P'_{\lambda_i}(\tilde{t}_i).$$

According to (21), for each $i = 1, \cdots, p$ we have

$$K_i = \begin{cases} \{(0,0)\} & \text{if } \big(\tilde{\beta}_i = 0 \text{ and } |\tilde{q}_i| < |P'_{\lambda_i}(\tilde{t}_i)|\big), \\ \{(\beta_i, t_i) \in \mathbb{R}^2_+ \mid \beta_i - t_i = 0\} & \text{if } \big(\tilde{\beta}_i = 0 \text{ and } \tilde{q}_i = -P'_{\lambda_i}(\tilde{t}_i)\big), \\ \{(\beta_i, t_i) \in \mathbb{R}^2 \mid \beta_i - t_i = 0\} & \text{if } \tilde{\beta}_i > 0, \\ \{(\beta_i, t_i) \in \mathbb{R}_- \times \mathbb{R}_+ \mid \beta_i + t_i = 0\} & \text{if } \big(\tilde{\beta}_i = 0 \text{ and } \tilde{q}_i = P'_{\lambda_i}(\tilde{t}_i)\big), \\ \{(\beta_i, t_i) \in \mathbb{R}^2 \mid \beta_i + t_i = 0\} & \text{if } \tilde{\beta}_i < 0, \end{cases} \tag{B.2}$$

and

$$K = \mathbb{R} \times K_1 \times \cdots \times K_p.$$

Next, we give an explicit expression for the affine hull of K. Define two matrices M and N as follows:

$$M = \begin{bmatrix} 1 & 0 \\ 0 & I_p \\ 0 & I_p \end{bmatrix} \text{ and } N = \begin{bmatrix} 1 & 0 \\ 0 & I_p \\ 0 & -I_p \end{bmatrix}.$$

Construct a matrix Ξ by first adding the common first column of M and N and then adding the $(i+1)^{th}$ column of M (N) if the condition in the second or third (fourth or fifth) row of (B.2) is satisfied. Columns of Ξ form a basis of the affine hull of K. Note that $\Xi^T L \Xi = Q$, where Q is defined in Assumption 3. From Proposition 2.5 and Theorem 4.3 of Robinson (1992), L_K is a global homeomorphism. Under Assumption 1(b), it is easy to see that the partial derivative of f_0 at $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ is strong. An application of (Robinson, 1995, Theorem

3) implies that z_0 is a locally unique solution to (19), therefore $(\tilde{\beta}_0, \tilde{\beta}, \tilde{t})$ is a locally unique optimal solution to (9).

□

Proof of Lemma 3. The conclusion follows from an application of (Lu and Budhiraja, 2013, Theorem 4). We verify the assumptions of the latter theorem as follows. Assumption 1 in Lu and Budhiraja (2013) holds under Assumptions 1 and 2 of this paper according to equations (13) and (15). Moreover, Assumption 4(a) in Lu and Budhiraja (2013) is satisfied for the compact set $\mathcal{C}$ under Assumption 4(a) of this paper.

□

Proof of Theorem 1. From the proof of Lemma 3 we know that Assumption 1 in Lu and Budhiraja (2013) holds. According to Lemma 2, Assumption 2 in Lu and Budhiraja (2013) holds under Assumptions 1-3 of this paper. Furthermore, Assumption 4(a-b) of this paper guarantees Assumption 4 in Lu and Budhiraja (2013) to be satisfied. Consequently, conclusions in this theorem follow from (Lu and Budhiraja, 2013, Theorem 7).

□

Proof of Theorem 2. The convergence results for $d\Pi_S(z_N)$ and $d(f_N)_S(z_N)$ in Case I follow from the fact that $z_N \to z_0$ almost surely and the continuity of $d\Pi_S(\cdot)$ and $d(f_N)_S(\cdot)$. Moreover, we can prove the following result using similar arguments in the proof of Corollary 3.2 in Lu (2014): there exists a positive real number ϕ such that

$$\lim_{N\to\infty} \text{Prob}\left\{ \sup_{h\in\mathbb{R}^{2p+1}} \frac{\|\Phi_N(z_N)(h) - L_K(h)\|}{\|h\|} < \frac{\phi}{g(N)} \right\} = 1, \tag{B.3}$$

which implies that $\Phi_N(z_N)$ converges to L_K in probability.

□

Proof of Theorem 3. This theorem can be proved using the same arguments in the proof of Theorem 3 in Lu et al. (2017).

□

Proof of Lemma 4. To show (48), we use the equation (19) with (13) and (14). With $\lambda = 0$, by plugging (47) into (13), we have

$$\begin{aligned} z_0 =& (\beta_0^{true}, \beta^{true}, t^{true}) - f_0(\beta_0^{true}, \beta^{true}, t^{true}) \\ =& \begin{bmatrix} \beta_0^{true} + 2E(Y - \beta_0^{true} - X^T\beta^{true}) \\ \beta^{true} + 2E[(Y - \beta_0^{true} - X^T\beta^{true})X] + 2m\beta^{true} \\ t^{true} - 2mt^{true} \end{bmatrix} = \begin{bmatrix} \beta_0^{true} \\ (1+2m)\beta^{true} \\ (1-2m)t^{true} \end{bmatrix}. \end{aligned} \tag{B.4}$$

Rearranging (B.4) proves (48).

□

Proof of Theorem 4. Recall that $(\beta_0^{true}, \beta^{true}, t^{true})$ and $(\hat{\beta}_0, \hat{\beta}, \hat{t})$ are solutions to

$$-f_0(\beta_0, \beta, t) \in N_S(\beta_0, \beta, t) \quad \text{and} \quad -f_N(\beta_0, \beta, t) \in N_S(\beta_0, \beta, t) \tag{B.5}$$

respectively, where

$$f_0(\beta_0, \beta, t) = \begin{bmatrix} -2E[Y - \beta_0 - \sum_{i=1}^p \beta_i X_i] \\ -2E[(Y - \beta_0 - \sum_{i=1}^p \beta_i X_i)X] - 2m\beta \\ 2mt \end{bmatrix}. \tag{B.6}$$

and

$$f_N(\beta_0, \beta, t) = \begin{bmatrix} -2N^{-1} \sum_{i=1}^N [y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij}] \\ -2N^{-1} \sum_{i=1}^N [(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})\mathbf{x}_i] - 2m\beta \\ (P'_{\lambda_i}(t_i) + 2mt_i)_{i=1}^p \end{bmatrix}. \tag{B.7}$$

By Assumption 1'(a-b), f_N almost surely converges to f_0 in the space of continuously differentiable functions on a neighborhood of $(\beta_0^{true}, \beta^{true}, t^{true})$. Moreover, by the functional central limit theorem, the first $p+1$ component functions of $\sqrt{N}(f_N - f_0)$ weakly converge to the random function $Y : \mathbb{R}^{2p+1} \to \mathbb{R}^{p+1}$, with $Y(\beta_0^{true}, \beta^{true}, t^{true}) \sim \mathcal{N}(0, \Sigma_0^{*1})$. By Assumption 1'(c) and the fact that $\lim_{N\to\infty} \sqrt{N}\lambda_i = c_i$, the last p component functions of $\sqrt{N}(f_N - f_0)$ converge to $(h_i)_{i=1}^p = \left(c_i \frac{\partial^2 P}{\partial\lambda_i \partial t_i}(0, t_i^{true})\right)_{i=1}^p$.

By the choice of m, the matrix L^* defined in (50) is positive definite. This implies that the normal map $L^*_{K^*}$ is a global homeomorphism. By (Lu and Budhiraja, 2013, Lemma 1), there exists a neighborhood of f_0 such that when f_N belongs to that neighborhood the solutions $(\hat{\beta}_0, \hat{\beta}, \hat{t})$ and z_N are well defined. We can then proceed similarly to the proof of (Lu and Budhiraja, 2013, Theorem 7) to show that

$$\sqrt{N}(G^*(z_N) - G^*(z_0^*)) \Rightarrow G^* \circ (L^*_{K^*})^{-1}(\mathcal{N}(0, \Sigma_0^{*1}), h),$$

which is (54).

□

Proof of Theorem 5. Consider the case where $h_i = 0$ for each i. Note that $\Sigma_0^* = \begin{bmatrix} \Sigma_0^{*1} & 0 \\ 0 & 0 \end{bmatrix}$.

Let $q_0 \in \mathbb{R}$ and $q \in \mathbb{R}^p$. We will simplify the expression of $G^* \circ (L^*_{K^*})^{-1}(q_0, q, 0)$. Consider the minimization problem

$$\min_{(\beta_0, \beta, t) \in K} \beta_0^2 + \beta^T(\Sigma - mI_p)\beta + \sum_{i=1}^p mt_i^2 - q_0\beta_0 - q^T\beta,$$

whose solution satisfies $(q_0, q, 0) \in L^*(\beta_0, \beta, t) + N_{K^*}(\beta_0, \beta, t)$. By the expression of K^* in (53), the above problem can be reduced to

$$\min_{(\beta_0, \beta) \in \mathbb{R}^{p+1}} \beta_0^2 + \beta^T\Sigma\beta - q_0\beta_0 - q^T\beta,$$

whose solution is given by $\beta_0 = \frac{1}{2}q_0$ and $\beta = \frac{1}{2}\Sigma^{-1}q$. The first $p+1$ components of $(L^*_{K^*})^{-1}(q_0, q, 0)$ are given by $(\frac{1}{2}q_0, (\frac{1}{2}+m)\Sigma^{-1}q)$, so

$$G^* \circ (L^*_{K^*})^{-1}(q_0, q, 0) = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2}\Sigma^{-1} \end{bmatrix} (q_0, q).$$

Furthermore, since Σ_0^{*1} is the covariance matrix of the first $p+1$ components of the random vector $F(\beta_0^{true}, \beta^{true}, t^{true}, X, Y)$, we can show that

$$\Sigma_0^{*1} = \begin{bmatrix} 4\sigma^2 & 0 \\ 0 & 4\sigma^2\Sigma \end{bmatrix}.$$

Therefore,

$$G^* \circ (L^*_{K^*})^{-1}(\mathcal{N}(0, \Sigma_0^{*1}), 0) = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2}\Sigma^{-1} \end{bmatrix} (\mathcal{N}(0, \Sigma_0^{*1})) = \mathcal{N}(0, \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2\Sigma^{-1} \end{bmatrix}).$$

Furthermore, by modifying the augments in the proof of Theorem 5 in Lu et al. (2017) via substituting H_N of this paper, we can show (55).

□

Below is a lemma that will be used in the proof of Theorem 6.

Lemma 5. *Suppose that Assumptions 1' (a-c), 2, and 4' (a-b) hold. Then $\hat{R}$ converges to R in probability uniformly on compact sets.*

Proof of Lemma 5. This lemma is similar to Lemma 5 in Lu et al. (2017) except for different expressions of R and $\hat{R}$. It can be proved using a similar argument. Let

$$T = (L^*_{K^*})^{-1} \begin{bmatrix} (\Sigma_0^{*1})^{\frac{1}{2}} & 0 \\ 0 & \mathrm{diag}(h_i)_{i=1}^p \end{bmatrix} \quad \text{and} \quad T_N = (\Phi_N(z_N))^{-1} \begin{bmatrix} (\Sigma_N^1)^{\frac{1}{2}} & 0 \\ 0 & \mathrm{diag}(\hat{h}_i)_{i=1}^p \end{bmatrix}.$$

Applying Proposition 2 in Lamm et al. (2014), we can check that T_N converges to T in probability uniformly on compact sets. Since G^* is a full rank matrix, we conclude that $\hat{R}$ converges to R in probability uniformly on compact sets.

□

Proof of Theorem 6. By Lemma 5, $\hat{R}_i$ converges to R_i in $C(\mathbb{R}^{2p+1}, \mathbb{R})$ in probability uniformly on compact sets. Let

$$Z_N = \sqrt{N}\left((\hat{\beta}_0^{true}, \hat{\beta}^{true}) - (\beta_0^{true}, \beta^{true})\right)_i$$

for $i = 1, \cdots, p+1$. From (54), Z_N converges to $R_i(Z)$ in distribution. Then the conclusions follow from Lemma 4 in Lu et al. (2017) with $u_N = \hat{R}_i$ and $u = R_i$.

□

C Example 5: Prostate cancer data

	Our method (GIC)		SVI-Lasso		LDPE		JM	
	Est	Ind CI	Est	Ind CI	Est	Ind CI	Est	Ind CI
β_1^{true}	0.73	[0.44, 1.01]	0.72	[0.42, 1.02]	0.70	[0.47, 0.93]	0.68	[0.03, 1.33]
β_2^{true}	0.28	[0.07, 0.49]	0.29	[0.09, 0.50]	0.28	[0.10, 0.46]	0.26	[-0.22, 0.75]
β_3^{true}	-0.08	[-0.31, 0.15]	-0.07	[-0.33, 0.18]	-0.09	[-0.29, 0.11]	-0.14	[-0.66, 0.38]
β_4^{true}	0.21	[-0.02, 0.45]	0.21	[-0.02, 0.45]	0.21	[0.01, 0.41]	0.21	[-0.31, 0.73]
β_5^{true}	0.33	[0.05, 0.60]	0.34	[0.04, 0.63]	0.31	[0.08, 0.54]	0.31	[-0.33, 0.94]
β_6^{true}	-0.20	[-0.47, 0.06]	-0.18	[-0.45, 0.09]	-0.20	[-0.47, 0.07]	-0.29	[-1.08, 0.50]
β_7^{true}	-0.05	[-0.30, 0.21]	-0.02	[-0.27, 0.24]	-0.01	[-0.27, 0.25]	-0.02	[-0.76, 0.72]
β_8^{true}	0.25	[-0.04, 0.54]	0.26	[-0.04, 0.56]	0.24	[-0.03, 0.51]	0.27	[-0.52, 1.05]

Table 4: Estimates and 95% individual CIs of true regression coefficients in the linear model for different methods computed from prostate cancer data.

In this real data example, we consider the prostate cancer dataset (Tibshirani, 1996) and compute the individual confidence intervals of the true regression coefficients with the confidence level 0.95. We use the same 67 training samples studied in Hastie et al. (2001). The data are standardized at the beginning of our analysis. For our proposed method, we use the MCP penalty with the parameter $a = 2$ and choose the best tuning parameter λ by GIC. Table 4 shows the estimates and confidence intervals of different parameters in the linear model. By checking whether each confidence interval contains zero or not, we can observe that our method and the SVI-Lasso method deliver the same inference results. However, compared with the SVI-Lasso method, the confidence intervals constructed by our proposed method are shorter in most cases. The results of our proposed method and the results of LDPE are also comparable. Compared with the other three methods, for this real data example, the confidence intervals constructed by the JM method are overall wider.

D Example: Inference of the population penalized parameter

Consider the following true linear model

$$Y = 2X_1 + X_2 + 3\epsilon,$$

where $\epsilon \sim \mathcal{N}(0, 1)$. The covariance matrix of $(X_1, X_2)^T$ is Σ where $\Sigma_{11} = \Sigma_{22} = 1$ and $\Sigma_{12} = \Sigma_{21} = 0.5$. If we use the LASSO penalty and choose $m = 0$, the objective function (9) in the manuscript is

$$\begin{aligned}
&\min_{\beta_0,\beta,t} (\beta^*-\beta)\Sigma(\beta^*-\beta)^T+\beta_0^2+\sigma^2+\lambda\sum_{i=1}^{p} t_i \\
&\text{s.t. } t_i-\beta_i \geqslant 0, \quad i=1,\cdots,p, \\
&\qquad t_i+\beta_i \geqslant 0, \quad i=1,\cdots,p,
\end{aligned}$$

where $\beta^*=(2,1)$ is the true parameter.

Suppose that $\lambda \leq 3$. Let $(\tilde\beta_0,\tilde\beta_1,\tilde\beta_2,\tilde t_1,\tilde t_2)=(0,2-\frac{\lambda}{3},1-\frac{\lambda}{3},2-\frac{\lambda}{3},1-\frac{\lambda}{3})$. We can check that $f_0(\tilde\beta_0,\tilde\beta_1,\tilde\beta_2,\tilde t_1,\tilde t_2)=(0,-\lambda,-\lambda,\lambda,\lambda)^T$. In addition, we have

$$N_S(\tilde\beta_0,\tilde\beta_1,\tilde\beta_2,\tilde t_1,\tilde t_2)=N_R(\tilde\beta_0)\times N_{S_1}(\tilde\beta_1,\tilde t_1)\times N_{S_2}(\tilde\beta_2,\tilde t_2),$$

where

$$\begin{aligned}
S_1&=\{(\beta_1,t_1)|t_1-\beta_1\geq 0, t_1+\beta_1\geq 0\}\\
S_2&=\{(\beta_2,t_2)|t_2-\beta_2\geq 0, t_2+\beta_2\geq 0\}.
\end{aligned}$$

Furthermore, we can check that

$$N_R(\tilde\beta_0)=\{0\},\ \ N_{S_1}(\tilde\beta_1,\tilde t_1)=N_{S_2}(\tilde\beta_2,\tilde t_2)=\{(v_1,v_2)\in R^2|v_1=-v_2\}.$$

Therefore, we have

$$-f_0(\tilde\beta_0,\tilde\beta_1,\tilde\beta_2,\tilde t_1,\tilde t_2)=(0,\lambda,\lambda,-\lambda,-\lambda)^T\in N_S(\tilde\beta_0,\tilde\beta_1,\tilde\beta_2,\tilde t_1,\tilde t_2).$$

So $(\tilde\beta_0,\tilde\beta_1,\tilde\beta_2,\tilde t_1,\tilde t_2)=(0,2-\frac{\lambda}{3},1-\frac{\lambda}{3},2-\frac{\lambda}{3},1-\frac{\lambda}{3})$ satisfies the variational inequality (17) and

$$\begin{aligned}
z_0=&(\tilde\beta_0,\tilde\beta_1,\tilde\beta_2,\tilde t_1,\tilde t_2)-f_0(\tilde\beta_0,\tilde\beta_1,\tilde\beta_2,\tilde t_1,\tilde t_2)\\
=&(0,2+\frac{2\lambda}{3},1+\frac{2\lambda}{3},2-\frac{4\lambda}{3},1-\frac{4\lambda}{3}).
\end{aligned}$$

If $\lambda=3$, we can check that $(z_0)_3=-(z_0)_5=3$, $((z_0)_3,(z_0)_5)\in C_2^4$, and therefore Π_K, G and $(L_K)^{-1}$ are all piecewise linear functions. In this case, the asymptotical distribution of the LASSO estimates $(\hat\beta_0,\hat\beta_1,\hat\beta_2)^T$ is non-normal.

However, if $0\leq\lambda<3$, we can check that $((z_0)_2,(z_0)_4)\notin C_1^3\cup C_1^4$ and $((z_0)_3,(z_0)_5)\notin C_2^3\cup C_2^4$, and therefore Π_K, G and $(L_K)^{-1}$ are all linear functions. We can check that

$$\Pi_K(h)=\begin{bmatrix}1&0&0&0&0\\0&1/2&0&1/2&0\\0&0&1/2&0&1/2\\0&1/2&0&1/2&0\\0&0&1/2&0&1/2\end{bmatrix}\begin{bmatrix}h_0\\h_1\\h_2\\h_3\\h_4\end{bmatrix},\ \text{where } h=(h_0,h_1,h_2,h_3,h_4)^T\in R^5,$$

$$L_K^{-1}(h) = \begin{bmatrix} 1/2 & 0 & 0 & 0 & 0 \\ 0 & 2/3 & -1/3 & -1/3 & -1/3 \\ 0 & -1/3 & 2/3 & -1/3 & -1/3 \\ 0 & 2/3 & -1/3 & 5/3 & -1/3 \\ 0 & -1/3 & 2/3 & -1/3 & 5/3 \end{bmatrix} \begin{bmatrix} h_0 \\ h_1 \\ h_2 \\ h_3 \\ h_4 \end{bmatrix}, \text{ where } h = (h_0, h_1, h_2, h_3, h_4)^T \in R^5,$$

and

$$\Sigma_0 = \mathrm{Cov}(F(\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2, \tilde{t}_1, \tilde{t}_2, X, Y)) = \begin{bmatrix} 36 + 4\lambda^2/3 & 0 & 0 & 0 & 0 \\ 0 & 36 + 7\lambda^2/3 & 18 + 5\lambda^2/3 & 0 & 0 \\ 0 & 18 + 5\lambda^2/3 & 36 + 7\lambda^2/3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

In addition, by Theorem 1, we have

$$\sqrt{N} L_K(z_N - z_0) \Rightarrow \mathcal{N}(0, \Sigma_0).$$

Therefore,

$$\sqrt{N}((\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{t}_1, \hat{t}_2)^T - (\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2, \tilde{t}_1, \tilde{t}_2)^T) \Rightarrow \Pi_K \circ (L_K)^{-1} \mathcal{N}(0, \Sigma_0).$$

By plugging in $\Pi_K, (L_K)^{-1}$ and Σ_0, we have

$$\sqrt{N}\left(\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} - \begin{bmatrix} \tilde{\beta}_0 \\ \tilde{\beta}_1 \\ \tilde{\beta}_2 \end{bmatrix} \right) \Rightarrow \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 9 + \lambda^2/3 & 0 & 0 \\ 0 & 12 + 5\lambda^2/9 & -6 - \lambda^2/9 \\ 0 & -6 - \lambda^2/9 & 12 + 5\lambda^2/9 \end{bmatrix} \right),$$

where $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ are the LASSO estimates and $\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2$ are the population penalized parameters. When $\lambda = 0$, the limiting distribution is the same as the distribution of the least squares estimator.

References

Facchinei, F. and Pang, J. S. (2003), *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Springer Series in Operations Research, New York: Springer.

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, New York: Springer.

Lamm, M., Lu, S., and Budhiraja, A. (2014), "Individual confidence intervals for true solutions to stochastic variational inequalities," *Submitted*.

Lu, S. (2014), "A new method to build confidence regions for solutions of stochastic variational inequalities," *Optimization: A Journal of Mathematical Programming and Operations Research*, 63, 1431–1443.

Lu, S. and Budhiraja, A. (2013), "Confidence regions for stochastic variational ienqualities," *Mathematics of Operations Research*, 38, 545–568.

Lu, S., Liu, Y., Yin, L., and Zhang, K. (2017), "Confidence intervals and regions for the lasso by using stochastic variational inequality techniques in optimization," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79, 589–611.

Robinson, S. M. (1992), "Normal maps induced by linear transformations," *Mathematics of Operations Research*, 17, 691–714.

— (1995), "Sensitivity analysis of variational inequalities by normal-map techniques," in *Variational Inequalities and Network Equilibrium Problems*, ed. Giannessi, F. and Maugeri, A., New York: Plenum Press, pp. 257–269.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B*, 58, 267–288.